CHAIR OF
ECONOMETRICS
FREIE UNIVERSITÄT BERLIN

Bachelor Thesis

Principal Component Analysis applied to the Term Structure: an analysis on European interest rate yields between 2004 to 2020

Ana Vitoria Rodrigues Lima

Supervisor:	PhD Elias Wolf, Prof. Dr. Dieter Nautz
Semester:	Summer Term 2020
Author:	Ana Vitoria Rodrigues Lima
Matriculation-Nr.:	5247230
Address:	Via Fontana 19/1, Pordenone 33080, Italy
Email:	vito.rodrigueslima@gmail.com
Telephone:	+49 (0)157 309 58 797
Subject:	B.Sc Economics (Volkswirtschaftslehre)
Submission:	10 November 2020

Abstract

The first three principal components (PCs) that result from a principal component analysis (PCA) of a term structure have been across past literature interpreted as shift (or level), slope (or steepness) and curvature of the yield curve (abbreviated as SSC). Furthermore, recent literature has argued that PCA applied in a negative interest rate environment is less effective compared to its application to a positive interest rate environment. This is relevant because this might affect the efficacy of PCA as a risk-management instrument in a potential bonds portfolio.

This bachelor thesis first dives into the technicalities of PCA and PCA applied to the term structure, to then empirically show the SSC pattern and to lastly re-examine whether the negativity of interest rates impacts unfavourably the efficacy (i.e. lowers the amount of variability explained) of the first three factors compared to a positive interest rate environment. No difference in efficacy, contrarily to recent literature, is found. Implicitly, this suggest that the amount of variability explained by the first 3 PCs might be affected by other idiosyncratic characteristics of a given sample.

Contents

List of Figures

The source of all the Figures comes from me, the Author's calculation, unless stated otherwise. For Figures 4.2, 4.3 and 4.4 the Python language and the library Plotly have been used. For the rest of the figures the language R and the outputs from R have been used.

List of Tables

The source of all the inputs of the Tables comes from me, the Author's calculation, unless stated otherwise. The language used to perform the data analysis in this paper has mainly been R. Python has been used as well, for Table 4.2.

Chapter 1

Introduction

This thesis focuses on the usage of a method of dimensionality reduction, principal component analysis (PCA), applied to the european yield curve. The dataset used here is the average of AAA-rated bonds in the Euro Area.

The dataset comprises time series starting from 6th of September 2004 to the 23rd of September 2020. For argumentation and demonstration purposes of this paper, this dataset will be divided in 5 parts: Part 1 comprises the entirety of the dataset, Part 2 comprises the time series between 2004 and 2007 within a positive interest rate environment, Part 3 comprises the time series between 2014 and 2017 within a negative interest rate environment, Part 4 comprises the time series between 2017 and 2020 within an even deeper negative interest rate environment, Part 5 is an extension of Part 3 with a time series between 2014 and 2020.

Previous literature (Steeley, 1990; Litterman & Scheinkman, 1991; Lord & Pelsser, 2006; Salinelli & Sgarra, 2007), has shown that the first three principal components (PCs) represent the shift, slope and curvature (SSC) of the yield curve and how these three components alone can explain the majority of variability of the entire yield curve. This paper will illustrate this through the Part 1 Dataset, as the first PC highly correlates with the shift, the second with the slope and the third with the curvature.

Furthermore, recent literature (Lazarevic, 2019) has shown that in a negative interest rate environment the amount of variability explained by the first three principal components is lower than the amount explained by the first three PCs in a positive interest rate environment. An implicit conclusion from Lazarevic (2019) is that the application of PCA to the term structure requires awareness of the business cycle of the underlying market. This paper will replicate the results from Lazarevic (2019) with the Part 2 and 3 Datasets.

This thesis paper expands this analysis with the Part 4 and 5 Datasets. Thereby, the expectation is that the variability explained by the first 3 PCs from the Part 4 and 5 Datasets will be lower than the one resulting from the Part 2 Dataset. This expectation is given by the fact that both Part 4 and 5 Datasets lie in deeply negative interest rate environments.

This paper organises as follows. In chapter 2, I will delve into the previous literature that dealt with the SSC pattern and the positive vs negative interest rate application of PCA. In chapter 3, I will briefly go into the technicalities of PCA with some examples of how this is used in practice in the field of finance when applied to the yield curve. In chapter 4, I will delve into: the data selected, how this has been pre-prepared, to finally arrive at the characteristics of its history. In chapter 5, I will illustrate my empirical results. In chapter 6, I will conclude.

Chapter 2

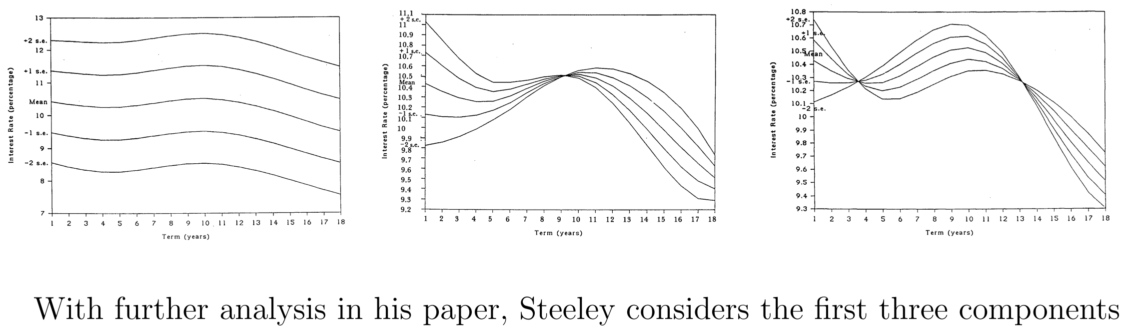
Literature Review

2.1 Shift, Slope and Curvature

The studies that pioneered analysing the shift, slope and curvature (SSC) pattern are the studies from Steeley (1990) and of Litterman & Scheinkman (1991). The former applied factor analysis on the UK bonds curve, known as “Gilts”. The latter applied factor analysis on the US bonds curve, known as “Treasuries”.

Steeley (1990) argues that using the price of coupon-bearing bonds would lead to a misleading price, because the coupon payment feature makes the interest rates (pure discount prices) ‘not directly observable’. As a consequence, he considers ‘pure’ discount prices, i.e. the ‘zeros’, which creates the spot curve[1]. First, he estimates B-spline coefficients, on which at first glance, he applies factor analysis. Given the striking and satisfactory results coming from this, he later applies factor analysis on the spot Gilts curve to “[...]obtain the sensitivities of the term structure to the changes (i.e. eigenvectors) without further transformations.” (Steeley, 1990, page 344). He considers the eigenvectors, coming from the latter analysis, to be the sensitivities of the spot rates to the factors. He then impacts the average spot rate curve by -2 and $+2$ times the eigenvectors for each of the three factors. A recurring characteristic of the SSC pattern is the number of sign changes. In the first PC there are 0 sign changes, in the second there is 1 sign change and in the third PC there are 2 sign changes. This results in the graphs of:

Figure 2.1: Impacts of Components 1, 2 and 3 of the Mean Term Structure[2]



With further analysis in his paper, Steeley considers the first three components to have the possibility to be interpreted as a change in level (shift), a change in

[1]See Chapter 4.1, where Spot, Par and Forward curves are explained.
[2]Source of these images: Steeley (1991), pages 346, 347 and 348.

slope and a change in curvature. At the end, as for the level, he assumes this to be possibly represented by a long rate (a maturity from the long end of the curve); he assumes the slope to be possibly represented by the spread rate (a long maturity minus a short maturity), and he assumes the curvature to be possibly represented by the volatility of the spread rate.

In Litterman & Scheinkman (1991) factor analysis is applied to the implied zero curve (spot curve) of the US bonds, the 'Treasuries'. This analysis is performed in the framework of hedging a bond portfolio. They refer to the first three factors to be 'level, steepness and curvature'. They consider factor analysis on the term structure as a useful tool to assess what a bond portfolio is exposed to, i.e. which kind of risks it is exposed to. In their analysis, the first three factors explain at least 96% of the variability of the excess returns of any zero (i.e. zero coupon bonds, spot curve). Given this, the effects to a bond portfolio could be reduced to three factors: level, steepness and curvature (to which also they, similarly to Steeley, refer to as changes in the rate volatility). This means that duration hedging alone cannot optimally hedge a portfolio, as the given portfolio would be immune to changes of the first factor only, while still being nonexempt to changes in the second and third factors.

Later, in Lekkos (2000) a critique to both of these studies is presented. This critique is about the choice, that Steeley and Litterman & Scheinkmann did in their respective papers, of using the spot curve - this choice is deemeed by Lekkos to be inappropriate as the spot rates are the averages of the corresponding forward rates due to the no-arbitrage condition[3]. Lekkos not only criticizes the usage of the spot curve instead of the forward curve, but also discredits the results of PCA applied to forward rates, which he considers not sufficient to draw a link between macroeconomic characteristics and the change of the shape of the yield curve that could be drawn from factor analysis (Lekkos, 2000, page 1).

The critique presented in Lekkos (2000) has been discussed in Lord & Pelsser (2006) to be incorrect, as the lack of the SSC pattern in Lekkos' results might be due to his non-smooth forward curve. Further, in light of the concepts of total positivity, oscillation matrices[4], of Green's matrix, and sign regularity, Lord and Pelsser formulate sufficient, although not necessary, properties for a correlation matrix to display the level and the slope[5]. However, with these properties the presence of curvature still remains unexplained. Nonetheless, the pattern of sign changes (0 for the first factor, 1 for the second factor and 2 for the third factor), although is not a full description of what is perceived as SSC, it is a recurrent pattern. Hence, given that "[...] in all the empirical studies [they] have seen, [this sign pattern] correctly signals the presence of level, slope and curvature, [...] [they] expect it to be sufficient"(Lord

[3]In Chapter 4 Section 1 the Forward vs Spot curve debate is covered in more detail.

[4]Oscillation matrices are a sub-class of total positive matrices. A property of oscillation matrices is that the n^{th} eigenvector of such a matrix has exactly $n-1$ sign changes (Lord & Pelsser, 2006).

[5]Lord & Pelsser (2006) - Page 21: "A quasi-correlation matrix **R** with strictly positive entries, that satisfies:
i) $\rho_{i,j+1} \leq \rho_{ij}$ for $j \geq i$, i.e. correlations decrease when we move away from the diagonal;
ii) $\rho_{i,j-1} \leq \rho_{ij}$ for $j \leq i$, same as $i)$
iii) $\rho_{i,i+j} \leq \rho_{i+1,i+j+1}$, i.e. the correlations increase when we move from northwest to southeast.
displays level and slope. "

& Pelsser, 2006, page 10). In their paper, Lord and Pelsser conclude that the part of PCA applied to the term structure that is an 'artefact' are the orthogonality of the factors and the extent to which the input is smooth enough as a curve. On the other hand, the part of PCA applied to the term structure that is a 'fact' are the positive correlations present in the term structures.

2.2 Positive Interest Rates Environment vs Negative Interest Rates Environment

Milan Lazarevic's paper from 2019 "Principal component analysis in the negative interest rate environment" has been the first paper in the space of PCA related to the yield curve to have assessed the difference in efficacy (amount of variability explained) between a positive and a negative interest rate environment.

Lazarevic applies principal component analysis to the average of AAA-rated bonds in the european area. His intention is to compare the results he gets from a positive interest rate environment sample (he picks a time-frame between 2004 and 2007) with the results coming from a negative interest rates environment sample (time-frame of 2014 to 2017). The aim for this is to check whether in the negative interest rates environment the established patterns of SSC continues to take place. Inspite of the inconsistencies[6] that arise from the correlation matrix of the negative interest rate dataset (replicated in this paper with the correlation matrix of Part 3 Dataset in Table 3.3), this correlation matrix still has the property of oscillation matrices (i.e. that the n^{th} eigenvector has $n-1$ sign changes) that provides the determined characteristics of the eigenvector influence and hence the SSC pattern (Lazarevic, 2019).

However, although the SSC pattern continues to take place in the negative interest rate environment, Lazarevic's results do not display the same amount of variability explained by the first three PCs from the positive rates sample and the negative rates sample. In the positive interest rates sample the first three PCs explain more than 95% of the entire variability, actually about 97%. This is not the case in the negative interest rates sample, where the first three PCs explain less than 95%. In order to reach the same amount of variance explained as in the positive rates sample, the consideration of an additional PC, the fourth PC, is needed in the negative interest rates sample.

Given the number of its sign changes (3 sign changes), the fourth factor is called by Lazarevic as "oscillatority" - this would need to be included in his negative interest rate environment assessment so that to reach the same amount of variability explained by the first 3 PCs in the positive interest environment sample. This leads to a request of additional attention in the case that PCA is used for risk management of a bond portfolio in a negative interest rate environment, as possible hedges that consider the first three factors would be less effective compared to a normal interest rates environment. Specifically, this implies that the application of PCA requires awareness of the business cycle of the underlying market.

[6] These inconsistencies are explained in Chapter 3.1.

Chapter 3

Methodology

3.1 PCA at a glance

Principal component analysis is a dimensionality reduction method whose purpose is to reduce the dimension of a large set of variables. This is performed while retaining as much of the variation as possible that is present in the original dataset, and hence of its information potential. Principal components (PCs) are uncorrelated and are calculated so that each of them maximises their variation under some given constraints that will be presented later. PCs are built so that the first principal component (PC) is the one whose data has the most variation, then the second PC has the second highest variation, and so on - PCs are constructed through calculating the covariance matrix of the original data and performing eigenvalue spectral decomposition on the covariance matrix, which makes the PCs orthogonal (statistically independent).

The first PC maximises its variance under the constraint that the sum of squared values in the first eigenvector is 1. The second PC maximises its variance under the constraint that the sum of squared values in the second eigenvector is 1 and that the covariance between the first PC and the second PC is 0. For the covariance between PCs to be 0, eigenvectors have to be orthogonal.[1] Meaning they have to be perpendicular (they intersect at an angle of 90°) or that their dot product (or more generally, an inner product) is 0. Orthogonality is necessary here also because the second PC should capture the highest variance from what is left after the first PC explains the data as much as it can. The third PC will be calculated in the same fashion as the second and so on.

PCA is performed by getting the dataset into the format of a matrix X. After this is done, the covariance matrix of the matrix X is taken into consideration and decomposed under what is known as spectral decomposition. However, the correlation matrix can also be used. Using the correlation matrix instead of the covariance matrix is equivalent to standardizing all the variables of the original dataset matrix X to have a mean of 0 and variance of 1. This paper applies PCA to the correlation matrices, these are presented in Tables from 3.1 to 3.5.

[1]See Appendix A.1 Derivation of the PCs and the mathematics behind this.

Furthermore, in the interest rates context, using the correlation matrix is interesting also in order to look at its features. As it is described by (Salinelli & Sgarra, 2006, page 683), the features of the correlation matrix of the forward interest rates are:

> (a) interest rates at different maturities are always positively correlated;
> (b) the correlation coefficients decrease when the distance between the indices increases: this is a far obvious consequence of the decreasing degree of correlation when the variables are more distant in time;
> (c) the previous reduction in the correlation between variables corresponding to the same difference in the indices tends to decrease as the maturities of both the variables are greater.

In Tables 3.1, 3.3, 3.5 there are some inconsistencies with the Salinelli - Sgarra aforementioned features. In Table 3.1 the inconsistencies are that $\rho_{1y15y,1y20y} > \rho_{1y10y}$, in Table 3.3 are that $\rho_{1y20y,1y30y} > \rho_{1y15y}$ and in Table 3.5 are that $\rho_{1y20y} > \rho_{1y15y}$. These inconsistencies, which Lazarevic (2019) refers to as anomalies, are not present in the correlation matrix, Table 3.2, of the Part 2 Dataset (which Lazarevic selected as the positive interest rates dataset). As a consequence, these inconsistencies are explained by him as coming from the markets distortions that negative interest rates caused. However, although the Part 4 Dataset is more negative than the Part 3 Dataset, these inconsistencies do not take place in its correlation matrix, see Table 3.4. Whether or not there is some idiosyncrasy found within the Part 3 Dataset, will be discussed later in this paper.

Table 3.1: Correlation Matrix - Part 1 Dataset [2]

Yields	$1y$	$2y$	$3y$	$4y$	$5y$	$7y$	$10y$	$15y$	$20y$	$30y$
$1y$	1	0.8706722	0.7382580	0.6541856	0.5864366	0.4944913	0.4292606	0.4581945	0.4929077	0.4263754
$2y$	0.8706722	1	0.9367560	0.8332568	0.7308554	0.5828366	0.4687370	0.4876800	0.5403280	0.4814921
$3y$	0.7382580	0.9367560	1	0.9571461	0.8832763	0.7341091	0.5671174	0.5397008	0.6039285	0.5612948
$4y$	0.6541856	0.8332568	0.9571461	1	0.9718646	0.8646631	0.6798993	0.6033408	0.6561303	0.6124999
$5y$	0.5864366	0.7308554	0.8832763	0.9718646	1	0.9430855	0.7754158	0.6621698	0.6823133	0.6171552
$7y$	0.4944913	0.5828366	0.7341091	0.8646631	0.9430855	1	0.9148106	0.7738225	0.7010445	0.5536456
$10y$	0.4292606	0.4687370	0.5671174	0.6798993	0.7754158	0.9148106	1	0.9119636	0.7298209	0.4442410
$15y$	**0.4581945**	0.4876800	0.5397008	0.6033408	0.6621698	0.7738225	0.9119636	1	0.8908696	0.5892942
$20y$	**0.4929077**	0.5403280	0.6039285	0.6561303	0.6823133	0.7010445	0.7298209	0.8908696	1	0.8743291
$30y$	0.4263754	0.4814921	0.5612948	0.6124999	0.6171552	0.5536456	0.4442410	0.5892942	0.8743291	1

Table 3.2: Correlation Matrix - Part 2 Dataset

Yields	$1y$	$2y$	$3y$	$4y$	$5y$	$7y$	$10y$	$15y$	$20y$	$30y$
$1y$	1	0.9367176	0.8429880	0.7317580	0.6241783	0.4916326	0.4432297	0.4151656	0.3997074	0.3854657
$2y$	0.9367176	1	0.9503044	0.8540818	0.7427311	0.5922157	0.5317987	0.5009702	0.4855831	0.4740138
$3y$	0.8429880	0.9503044	1	0.9562913	0.8844120	0.7458737	0.6402225	0.5468252	0.5095540	0.4943103
$4y$	0.7317580	0.8540818	0.9562913	1	0.9668195	0.8678876	0.7337163	0.5720302	0.5065002	0.4829740
$5y$	0.6241783	0.7427311	0.8844120	0.9668195	1	0.9405138	0.8072030	0.6041727	0.5152359	0.4831126
$7y$	0.4916326	0.5922157	0.7458737	0.8678876	0.9405138	1	0.9192545	0.7139286	0.6021257	0.5520158
$10y$	0.4432297	0.5317987	0.6402225	0.7337163	0.8072030	0.9192545	1	0.8924462	0.7999755	0.7475159
$15y$	0.4151656	0.5009702	0.5468252	0.5720302	0.6041727	0.7139286	0.8924462	1	0.9662231	0.9395292
$20y$	0.3997074	0.4855831	0.5095540	0.5065002	0.5152359	0.6021257	0.7999755	0.9662231	1	0.9799224
$30y$	0.3854657	0.4740138	0.4943103	0.4829740	0.4831126	0.5520158	0.7475159	0.9395292	0.9799224	1

[2] All the datasets have the first differences taken in order to ensure stationarity, unless stated otherwise. Hence, all the Tables from 3.1 to 3.5 also come from calculations coming from the dataset with the first differences taken. Additionally, these are the correlation matrices of the forward rates, not the spot rates. Refer to Chapter 4.1 for the explanation of this choice.

Table 3.3: Correlation Matrix - Part 3 Dataset

Yields	$1y$	$2y$	$3y$	$4y$	$5y$	$7y$	$10y$	$15y$	$20y$	$30y$
$1y$	1	0.7467732	0.5675521	0.4915096	0.4317059	0.3754605	0.3108292	0.2802612	0.3113723	0.3103705
$2y$	0.7467732	1	0.9033645	0.7816682	0.6680555	0.5505669	0.4814009	0.4533441	0.4294919	0.2841078
$3y$	0.5675521	0.9033645	1	0.9445068	0.8657556	0.7541648	0.6616158	0.6077071	0.5690253	0.3738293
$4y$	0.4915096	0.7816682	0.9445068	1	0.9660479	0.8920900	0.7884906	0.7105566	0.6811410	0.4851627
$5y$	0.4317059	0.6680555	0.8657556	0.9660479	1	0.9616110	0.8703719	0.7860346	0.7636917	0.5658801
$7y$	0.3754605	0.5505669	0.7541648	0.8920900	0.9616110	1	0.9527289	0.8808913	0.8482403	0.6052317
$10y$	0.3108292	0.4814009	0.6616158	0.7884906	0.8703719	0.9527289	1	0.9682210	0.8995250	0.5473875
$15y$	0.2802612	0.4533441	0.6077071	0.7105566	0.7860346	0.8808913	0.9682210	1	0.9415493	0.5585883
$20y$	**0.3113723**	0.4294919	0.5690253	0.6811410	0.7636917	0.8482403	0.8995250	0.9415493	1	0.7825732
$30y$	**0.3103705**	0.2841078	0.3738293	0.4851627	0.5658801	0.6052317	0.5473875	0.5585883	0.7825732	1

Table 3.4: Correlation Matrix - Part 4 Dataset

Yields	$1y$	$2y$	$3y$	$4y$	$5y$	$7y$	$10y$	$15y$	$20y$	$30y$
$1y$	1	0.8637758	0.7258025	0.6361242	0.5907675	0.5345038	0.5049394	0.4518187	0.4185470	0.3843984
$2y$	0.8637758	1	0.9415044	0.8752534	0.8167446	0.7170102	0.6286403	0.5628753	0.5348181	0.4951334
$3y$	0.7258025	0.9415044	1	0.9658624	0.9253162	0.8287636	0.7150937	0.6313999	0.6030497	0.5600448
$4y$	0.6361242	0.8752534	0.9658624	1	0.9734609	0.9026154	0.7881030	0.6837533	0.6453678	0.6015276
$5y$	0.5907675	0.8167446	0.9253162	0.9734609	1	0.9560526	0.8586238	0.7393778	0.6948859	0.6485815
$7y$	0.5345038	0.7170102	0.8287636	0.9026154	0.9560526	1	0.9486314	0.8406725	0.7899512	0.7411981
$10y$	0.5049394	0.6286403	0.7150937	0.7881030	0.8586238	0.9486314	1	0.9437246	0.9007358	0.8426645
$15y$	0.4518187	0.5628753	0.6313999	0.6837533	0.7393778	0.8406725	0.9437246	1	0.9809048	0.9224890
$20y$	0.4185470	0.5348181	0.6030497	0.6453678	0.6948859	0.7899512	0.9007358	0.9809048	1	0.9570918
$30y$	0.3843984	0.4951334	0.5600448	0.6015276	0.6485815	0.7411981	0.8426645	0.9224890	0.9570918	1

Table 3.5: Correlation Matrix - Part 5 Dataset

Yields	$1y$	$2y$	$3y$	$4y$	$5y$	$7y$	$10y$	$15y$	$20y$	$30y$
$1y$	1	0.8086258	0.6480536	0.5613011	0.5026965	0.4380749	0.3817116	0.3443029	0.3520006	0.3237803
$2y$	0.8086258	1	0.9221333	0.8252094	0.7316859	0.6127357	0.5287393	0.4878069	0.4666028	0.3478044
$3y$	0.6480536	0.9221333	1	0.9539584	0.8891575	0.7772547	0.6710669	0.6076928	0.5764638	0.4261923
$4y$	0.5613011	0.8252094	0.9539584	1	0.9681661	0.8912420	0.7796717	0.6934408	0.6626295	0.5124021
$5y$	0.5026965	0.7316859	0.8891575	0.9681661	1	0.9579499	0.8617840	0.7660175	0.7363767	0.5837084
$7y$	0.4380749	0.6127357	0.7772547	0.8912420	0.9579499	1	0.9505043	0.8669259	0.8278581	0.6408883
$10y$	0.3817116	0.5287393	0.6710669	0.7796717	0.8617840	0.9505043	1	0.9605522	0.8991456	0.6268919
$15y$	0.3443029	0.4878069	0.6076928	0.6934408	0.7660175	0.8669259	0.9605522	1	0.9540110	0.6579448
$20y$	**0.3520006**	0.4666028	0.5764638	0.6626295	0.7363767	0.8278581	0.8991456	0.9540110	1	0.8288959
$30y$	0.3237803	0.3478044	0.4261923	0.5124021	0.5837084	0.6408883	0.6268919	0.6579448	0.8288959	1

3.2 PCA applied to the yield curve

On a typical yield curve principal component analysis, the first factor explains more than 90% of the yield curve variation and the first three factors together even more so. Meaning that the whole information of an interest rates market (here the euro area, but this could also be applied to a specific country market, e.g. Spain or Germany) can be expressed just with three numbers, i.e. the first three PCs. Given that principal components, as previously mentioned, are orthogonal, they are uncorrelated and independent from each other; meaning that their respective information packages are uncorrelated. Hence, each of them can explain something different from each other. As a consequence, these factors can become quite useful in trade ideas if they can also be intuitively explained. In the context of interest rates, the first PC is explained as the shift, the second as the slope and the third as the curvature.

[3]The arbitrarity of the sign is to be noted. As stated in Joliffe (2002) at page 67: "It should be noted that the sign of any PC is completely arbitrary. If every coefficient in a PC, $z_k = \alpha_k' x$, has its sign reversed, the variance of z_k is unchanged, and so is the orthogonality of α_k with all other eigenvectors.". Given this, I set the sign so that the first PC always has positive signs. This arbitrary choice has been done for an economic reason. In case there is a negative shock, if the signs were negative it means that the yields would go up. Although this might be at first the initial reaction of the markets, we know that the Central Bank would also intervene with the decision to cut rates. Hence, the yield curve would rather shift downwards due to ECB rate cuts. As a consequence, the first PC1 has to be of a positive sign for these explanation purposes. With a

Table 3.6: PCA results from Part 1 Dataset [3]

Variables	$PC1$	$PC2$	$PC3$	$PC4$
$1y$	0.26967	0.422346	0.1681428	0.56076
$2y$	0.30679	0.455693	0.0814210	0.19191
$3y$	0.33502	0.338424	-0.0390552	-0.16647
$4y$	0.34984	0.183643	-0.1423547	-0.32499
$5y$	0.35111	0.040809	-0.2310796	-0.34277
$7y$	0.33817	-0.173813	-0.3599548	-0.16261
$10y$	0.30842	-0.357454	-0.3746351	0.25289
$15y$	0.30561	-0.408218	-0.0018212	0.41612
$20y$	0.31537	-0.326600	0.4145866	0.08241
$30y$	0.27005	-0.180976	0.6694075	-0.35715
Eigenvalues	7.15358	1.29855	0.78319	0.53229
Proportion of Variance	71.54%	12.98%	7.832%	5.323%
Cumulative Variance	71.54%	84.52%	92.353%	97.676%

To interpret the entries from Table 3.6: if the first eigenvector increases by 1 unit, all yields increase - meaning that the 1y yield would increase by 0.269, the 2y yield by 0.307, the 3y yield by 0.335 and so on. If the second eigenvector increases, then short yields increase and long yields decrease - the second PC explains the directional impact on the slope, which is not explained by the first factor.[4] If the third PC increases, then the front end of the curve increases, the belly decreases and the long end increases - this is interpreted as the curvature dynamics of the curve, which is not explained by the first nor by the second PC.

PCA applied to the term structure is a very useful instrument within finance. Given the aforementioned properties of PCA, we are able to explain a large dataset with few factors that not only explain the majority of its variability and retain a large part of the original information package of the original dataset, but that are also uncorrelated with each other, so that each of them 'explains something different'. To apply PCA to financial data however, there is also the assumption that the market is driven by a set of uncorrelated linear factors (Huggins & Schaller, 2013). If this is the case, this can be used as a powerful instrument to brainstorm trade ideas and as a risk-management instrument as well.

An example of PCA used to build trade ideas is to regress PCs to find the driving forces of these PCs - so that to then build a view on those driving forces. This would mean to use regressions to find external potential explaining variables for the given factors. Previous literature has already worked towards the attempt at giving an economical meaning to factors coming from principal component analysis - an example of this is Ponomareva, Sheen, & Wang (2019), who calculated PCs from bilateral $-FX pairs and researched potential economic meanings behind these through regressions in an auto-regressive model.

A further usage of PCA in the trading field is also to quantify the impact of these

positive sign in the first PC, it is possible to explain a negative and a positive shock without leading into contradictions. In case a negative shock hits the economy: the ECB intervenes, cuts rates, and the yield curve shifts down. In the case a positive shock hits the economy: the ECB would feel more comfortable adjusting the deposit interest rate at a higher rate to prevent a possible inflation spike. This is also confirmed by Ruppert and Matteson with "The first, o_1, has all positive values." (Ruppert & Matteson, 2015, page 522).

[4]Short yields are the maturities at the front of the curve - this is called the 'front end' of the curve as it has the short term maturities, e.g. 1y, 2y, 3y yields. Long yields are the maturities in the 'long end' of the curve, i.e. after for example the 20y and 30y yields. The belly of the yield curve is the maturities in between, such as 7y, 10y and 15y yields.

potential driving forces on the calculated factors and thereby hedge these specific factors accordingly. This can help in calculating optimal hedge ratios to immunize a portfolio against the changes in factors (Huggins & Schaller, 2013).

A further example of trading strategies that could deploy PCA are relative value trades. Relative value trades are trades defined as offering return opportunities that are uncorrelated to the market direction - more commonly these are also referred to as 'finding mispricings in the markets'. Consequently, PCA can be seen as a key in relative value analysis given its ability to produce and analyze time series uncorrelated with market direction (Huggins & Schaller, 2013).

Lastly, another example for PCA application in finance is the evaluation of the VaR, the Value at Risk. The aim of VaR calculations and assessments is to be able to formulate statemets such as "We are X per cent certain that we will not lose more than V dollars in the next N days." (Hull, 2015, page 517).

Chapter 4

Data

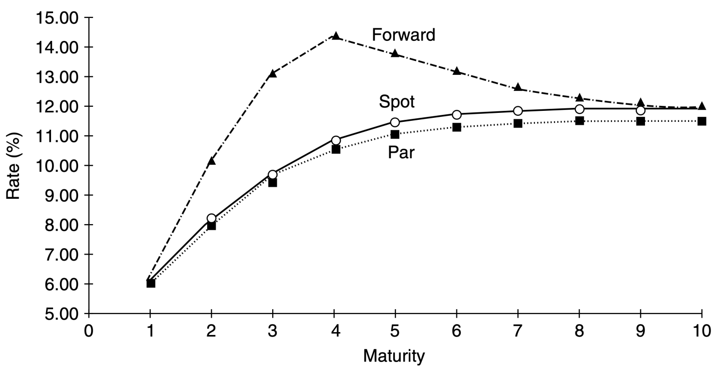
4.1 Data Selection and Elaboration

The dataset chosen comes from the public **Eurostat** Database. For this paper the average of AAA-rated euro area central government bonds has been selected for the maturities of 1y, 2y, 3y, 4y, 5y, 7y, 10y, 15y, 20y, and 30y[1] as daily data. The curve of reference is not the spot curve but the forward curve instead. The timeframe of the aforementioned dataset spans from the 6th of September 2004 to the 23rd of September 2020. The calculations shown in this paper have been done mainly in R[2] with just one step in Excel prior to that. Meaning that I downloaded the daily data of the forward curve for the AAA-rated euro area bonds in Excel, but before working on them in R, I calculated the first difference to ensure stationarity.

Forward curve instead of a spot curve

In this section I will go through why the forward curve has been selected instead of the spot curve for the analysis present in this paper. Before getting to this, I will briefly describe some characteristics of an yield curve and its different curves.

Figure 4.1: Par, Spot and One-Year Forward Rate Curves[3]



[1]These maturities have been chosen to replicate as similarly as possible the results from Lazarevic (2019) with the Part 2 and 3 Dataset in Chapter 5. Note that the Eurostat Dataset provides more maturities than this, i.e. all maturities ranging from 1 to 30 years.

[2]The R version used is 4.0.1 (2020-06-06). The additional packages ‘tseries’ and ‘stargazer’ as software have been used.

A yield curve is a curve plotted on a graph with the bond yields (usually plotted on the y axis) against their maturities (plotted on the x axis). The yield curve can be seen from different perspectives, among these it can be seem from the perspective of three curves: the par curve, the spot curve and the forward curve. The par curve is a curve constructed for theoretical bonds whose prices equal par (100) - note that at par the coupon equals the yield. The spot curve is the curve of the "zeros", i.e. the zero coupon paying bonds. The forward curve plots the forward rates, which are the interest rates for a loan between any two dates in the future as of today. Because of the no-arbitrage market assumption, a forward rate must be so that the interest rate of a n years loan is the same as if I had taken a loan today for n-i (or m) years with, still today, taken a loan in n-i (or m) years from now for i (or n-m) years.

$$(1 + s_n)^n = (1 + s_m)^m (1 + f_{m,n})^{n-m}$$

For example, if I were to take a loan today for 3 years with a 3y interest rate, I would have to pay the same as if I were to take two separate loans today: a loan today for 2 years with the 2y interest rate and to sign a contract today where I get a loan in 2 years from now for 1 year in length - the interest rate of the latter loan is the forward interest rate, 2 years forward for 1 year. To clarify this with a more straightforward numerical example:

$$(1 + s_3)^3 = (1 + s_2)^2 (1 + f_{2,3})^1$$

As Fabozzi (1983) summarizes: "[...] a par rate is used to discount a set of cash flows (those of a par bond) to today, a spot rate is used to discount a single future cash flow to today, and a forward rate is used to discount a single future cash flow to another (nearer) future date. The par yield curve, the spot-rate curve, and the forward-rate curve contain the same information about today's term structure of interest rates."[4]. Another perspective to look at the relationship among these curves, is that the 1y forward rates shows the 'marginal' yield of lengthening the maturity of the investment of one year - whereby the spot rates measure an investment's average reward from today to a determined maturity (Fabozzi, 1983, page 162). Given this, spot interest rates can be seen as geometric averages of one or more forward rates (Fabozzi, 1983). Because of this reason, Lekkos (2000) considers the results coming from the factor analysis applied to the term structure with the usage of spot curves to be a "[...] statistical artifact created by the restrictions that the no-arbitrate hypothesis imposes on the correlation matrix of spot interest rates".

In his piece from 2000, Lekkos considered the studies on the Gilts curve (Steeley, 1990) and on the Treasuries curve (Litterman & Scheinkman, 1991) to have arrived at a misinterpretation of the results. These studies used the 'zeros', i.e. the zero coupon paying bond yields on the spot curve. As stated above, the spot rates can be seen as a geometric average of the forward rates - hence Lekkos considered the spot rates to be a 'transformation' of the existing forward dates. This leads to an extra covariation to be artificially created, hence "[...] any incremental information extracted from spot rates that is not present in the forward rates has no economic

[3] Source of this image: Fabozzi (1983, page 162).

[4] These curves can be derived through interpolation and bootstrapping techniques - this will not be investigated in this paper. For this refer to Fabozzi (1983, page 179).

underpinnings and should be ignored." (Lekkos, 2000, page 8). A solution for that might seem at first to use the forward curve instead. However Lekkos's results of factor analysis applied to the forward yield curve are not satisfactory in determining the characteristics of shift, slope and curvature that the studies from Steeley and Litterman-Scheinkman showed. In Lekkos (2000), the factor analysis on forward rate reveals a very different factor structure, with only about 40 to 60% of explained variance on the first factor - hence, he concludes that there is "[...] no inference on the link between macroeconomic factors and changes of the shape of the term structure [that] can be drawn from the results of factor analysis." (Lekkos, 2000, page 1).

This has been later questioned by Lord & Pelsser (2006), where they do agree that the forward curve is a more appropriate curve to perform factor analysis on the term structure, but they consider the findings of Lekkos (2000) to have been misleading becasue of the technique Lekkos used to fit the forward curve. As Lekkos used a bootstrapping method to linearly interpolate between the quotes, his forward curve had 'kinks' instead of being a smooth enough yield curve. In their paper, Lord and Pelsser used other fitting techniques, such as continuously compounded annual forward rates - with which, contrarily to Lekkos (2000), they do find the level, slope and curvature pattern.

In light of this previous literature, the curve chosen for the analysis taking place in this paper is the forward curve.

Ensuring stationarity

A nonstationary model is a model where the volatility parameters are a function of time (Hull, 2015, page 832). If a time series is stationary, its properties do not depend on the time at which these series are observed. When doing statistical inference involving time series models, stationarity is a requirement.

To check if a process is an unit root process, one could compute and solve the lag polynomial. When the lag polynomial of a stochastic process has a root that is 1, this is called a unit root (or random walk) process. A process is stationary if the roots of a given lag polynomial are outside the unit circle. Alternatively, stationarity can be tested. In order to check if the given dataset is stationary, I tested the time series of the forward curve made of the average values of the AAA-rated 1y, 2y, 3y, 4y, 5y, 7y, 10y, 15y, 20y, 30y maturities bonds. The test used for this has been the augmented Dickey-Fuller test (ADF). The ADF test is an augmented version of the Dickey-Fuller test, meaning that the ADF is for more than 1 lag. The null hypothesis H_0 of the ADF test is that a unit root is present in a time series sample. The alternative hypothesis H_1 is stationarity or trend-stationarity. The aforementioned time series did not prove to be stationary with a significance level (alpha α) of nor 0.5 and nor 0.1 - this can be seen by comparing the p-values from Table 4.1. The p-values of the dataset without the first differences are mostly above an α of 0.1 - which leads to not being able to reject a unit root process. The rejection of the H_0 is however possible with the dataset where the first differences were taken, given that all p-values were 0.01, hence all smaller than an alpha of 0.1.

In light of these results, the decision of taking the first differences of all the datasets comes into place. Hence, all the datasets present in this paper, from Part 1 Dataset to Part 5 Dataset, have the first differences taken - unless stated otherwise in selected situations.

Table 4.1: Augmented Dickey Fuller Test

	$1y$	$2y$	$3y$	$4y$	$5y$	$7y$	$10y$	$15y$	$20y$	$30y$
Results from Part1 Dataset - without first differences taken										
$Dickey-Fuller$	-1.5117	-1.617	-1.7744	-1.9557	-2.1252	-2.327	-2.3436	-2.2387	-2.2579	-2.6229
$Lag\ order$	15	15	15	15	15	15	15	15	15	15
$p-value$	0.7851	0.7405	0.6738	0.5971	0.5253	0.4399	0.4327	0.4772	0.4691	0.3146
Results from Part1 Dataset - with first differences taken										
$Dickey-Fuller$	-15.449	-16.985	-16.854	-16.611	-16.499	-16.529	-16.906	-17.652	-18.652	-18.742
$Lag\ order$	15	15	15	15	15	15	15	15	15	15
$p-value$	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Further, in a stationary process, the auto-correlation function (ACF) decays in a very fast way. Whereas in a unit root process the ACF decays very slowly (and linearly), which makes it difficult to reject the H_0 in the ADF test[5].

Additionally, my decision to take the first differences of the given dataset of forward rates is in accordance with the previous literature on this topic, Lekkos (2000), Lord & Pelsser (2006) and Lazarevic (2019). Lekkos (2000) used the first differences as this reduces the correlation between interest rates and allows the study of the factors influencing interest rate movements.

4.2 Data Partition

As stated in Chapter 4 Section 1, the entire dataset selected comprises of a time series between 2004 and 2020 of the AAA-rated bonds from the European area - the curve taken into consideration is the forward curve and the selected points out of the yield curve are the 1y, 2y, 3y, 4y, 5y, 7y, 10y, 15y, 20y and 30y maturities. This paper is dividing this entire dataset into 5 parts for demonstration and argumentation purposes.

The first part, Part 1 Dataset, is the entire dataset. This starts on the 6th of September 2004 and ends on the 23rd of September 2020. The Part 1 Dataset will be used to demonstrate that the first three PCs correspond to the shift, slope and curvature of an yield curve in Chapter 5 Section 1.

The second part and the third part, Part 2 Dataset and Part 3 Dataset, have been selected as in Lazarevic (2019). This has been done because this paper first aims at replicating his results in Chapter 5 Section 2 - to thereafter expand them. Consequently, Part 2 has the same time series as the dataset characterised by the positive interest rate environment selected by Lazarevic - which in his paper from

[5]See Appendix B.1 for the ACF graphs of the time series with and without the first differences taken.

[6]The summary presented in Table 4.2, clearly, is the summary of all the original datasets - here the first differences are not taken.

Table 4.2: Summary of statistics of forward interest rates of AAA rating EMU zone member states [6]

Parameter	$1y$	$2y$	$3y$	$4y$	$5y$	$7y$	$10y$	$15y$	$20y$	$30y$
Part 1 Dataset										
$Count$	4088	4088	4088	4088	4088	4088	4088	4088	4088	4088
$Mean$	0.8060	0.9362	1.0903	1.2583	1.4289	1.7468	2.1194	2.4789	2.6353	2.7056
Std	1.5960	1.6224	1.6327	1.6362	1.6362	1.631506	1.6219	1.6026	1.576	21.5210
$Kurtosis$	-0.5174	-0.8051	-1.0692	-1.2590	-1.3821	-1.4930	-1.4994	-1.4259	-1.3722	-1.3245
$Skewness$	0.9178	0.7304	0.5483	0.3934	0.2643	0.0678	-0.1183	-0.2549	-0.2828	-0.2199
Min	-0.9100	-0.9700	-1.0000	-1.0100	-1.0000	-0.9400	-0.8200	-0.6300	-0.5100	-0.4300
Max	4.5400	4.7100	4.7400	4.730	4.7300	4.7400	4.7800	4.8700	4.9800	5.1800
25%	-0.5900	-0.5600	-0.4700	-0.3300	-0.1700	0.1000	0.4700	0.8600	1.0700	1.2800
50%	0.1000	0.2500	0.4700	0.7400	1.0000	1.5150	2.1300	2.6900	2.8700	2.8000
75%	2.0825	2.2125	2.4000	2.63250	2.8900	3.2700	3.6700	3.9800	4.1000	4.1100
Part 2 Dataset										
$Count$	853	853	853	853	853	853	853	853	853	853
$Mean$	3.0486	3.1778	3.2719	3.3619	3.4505	3.6138	3.8067	4.0075	4.1199	4.2353
Std	0.7787	0.7285	0.6556	0.5875	0.5294	0.4449	0.3767	0.3387	0.3323	0.3358
$Skewmess$	0.0625	0.0036	0.0119	0.0346	0.0522	0.0620	0.0459	0.0484	0.0912	0.1860
$Kurtosis$	-1.5693	-1.4420	-1.3112	-1.1838	-1.0673	-0.8994	-0.8581	-1.0262	-1.0956	-1.0333
Min	1.9300	2.0000	2.1500	2.3300	2.5100	2.7800	3.0600	3.3500	3.5000	3.5900
Max	4.3100	4.4700	4.5100	4.5400	4.5500	4.5900	4.6400	4.7100	4.7800	4.9900
25%	2.2200	2.4300	2.6500	2.8600	3.0000	3.2500	3.5000	3.7300	3.8200	3.9300
50%	3.0800	3.3100	3.4300	3.4900	3.5400	3.6400	3.8100	4.0100	4.1400	4.2400
75%	3.8300	3.8500	3.8500	3.8600	3.8800	3.9400	4.0700	4.2800	4.4300	4.5300
Part 3 Dataset										
$Count$	839	839	839	839	839	839	839	839	839	839
$Mean$	-0.4424	-0.4261	-0.3645	-0.2610	-0.1313	0.1509	0.5263	0.9357	1.1516	1.3188
Std	0.2665	0.2665	0.2699	0.2759	0.2866	0.3195	0.3808	0.4592	0.4844	0.4392
$Kurtosis$	-1.3708	-1.4233	-1.3762	-1.2123	-0.9488	-0.4204	-0.0240	0.1214	0.0536	-0.1967
$Skewness$	0.0827	0.0710	0.1676	0.2718	0.3458	0.4120	0.4877	0.5755	0.5313	0.1684
Min	-0.9100	-0.9300	-0.8700	-0.7300	-0.6100	-0.4300	-0.1700	0.1200	0.2800	0.4400
Max	0.0200	0.0400	0.1700	0.3400	0.5500	0.9900	1.5800	2.2300	2.4900	2.4300
25%	-0.7000	-0.6800	-0.6250	-0.5000	-0.3600	-0.0700	0.2900	0.6550	0.8550	1.0550
50%	-0.4500	-0.4600	-0.4300	-0.3300	-0.1600	0.0800	0.4500	0.8600	1.0800	1.3000
75%	-0.2400	-0.2000	-0.1400	-0.0300	0.1000	0.3600	0.7500	1.1900	1.4100	1.5600
Part 4 Dataset										
$Count$	944	944	944	944	944	944	944	944	944	944
$Mean$	-0.6909	-0.6744	-0.6073	-0.5083	-0.3938	-0.1608	0.1316	0.4441	0.6206	0.8020
Std	0.0736	0.0861	0.1269	0.1788	0.2313	0.3213	0.4117	0.4833	0.5150	0.5447
$Kurtosis$	-0.0399	0.5873	-0.0044	-0.4479	-0.7542	-1.1203	-1.3295	-1.4068	-1.4111	-1.3963
$Skewness$	-0.6752	-0.7810	-0.2083	-0.0808	-0.1366	-0.2833	-0.3872	-0.4392	-0.4574	-0.4698
Min	-0.9100	-0.9700	-1.0000	-1.0100	-1.0000	-0.9400	-0.8200	-0.6300	-0.5100	-0.4300
Max	-0.5600	-0.5100	-0.3100	-0.0900	0.1300	0.4800	0.8400	1.1500	1.3200	1.5000
25%	-0.7400	-0.7200	-0.6800	-0.6500	-0.5900	-0.4800	-0.3000	-0.0700	0.0700	0.2200
50%	-0.6800	-0.6600	-0.6100	-0.4900	-0.3500	-0.0600	0.3200	0.6900	0.8800	1.0700
75%	-0.6300	-0.6100	-0.5200	-0.3875	-0.2400	0.0800	0.4600	0.8600	1.0700	1.2800
Part 5 Dataset										
$Count$	1597	1597	1597	1597	1597	1597	1597	1597	1597	1597
$Mean$	-0.5531	-0.5374	-0.4758	-0.3792	-0.2628	-0.0165	0.3052	0.6564	0.8480	1.0166
Std	0.2304	0.2331	0.2460	0.2710	0.3040	0.3750	0.4661	0.5582	0.5938	0.5849
$Kurtosis$	-0.1875	-0.2857	-0.3414	-0.4183	-0.4432	-0.4133	-0.3114	-0.1662	-0.1939	-0.4844
$Skewness$	0.9910	0.9065	0.7254	0.4886	0.2832	0.0495	-0.0140	0.0542	0.0290	-0.2649
Min	-0.9100	-0.9700	-1.0000	-1.0100	-1.0000	-0.9400	-0.8200	-0.6300	-0.5100	-0.4300
Max	0.0200	0.0400	0.1700	0.3400	0.5500	0.9900	1.5800	2.2300	2.4900	2.4300
25%	-0.7100	-0.6900	-0.6500	-0.5900	-0.5000	-0.3200	-0.0600	0.2600	0.4400	0.6100
50%	-0.6400	-0.6200	-0.5400	-0.4300	-0.3000	0.0100	0.3700	0.7500	0.9500	1.1400
75%	-0.4200	-0.3900	-0.3200	-0.1800	-0.0500	0.2300	0.5900	0.9700	1.1700	1.3700

2019 he refers to as 'normal times'. Thence the Part 2 Dataset ranges from the 7th of September 2004 to the 31st of December 2007. Also the Part 3 Dataset has been chosen accordingly to Lazarevic's selection - the Part 3 Dataset corresponds to his chosen negative interest rates environment sample. This ranges from the 6th of June 2014 to the 21st of September 2017.

The fourth and fifth partitions are the datasets used to perform the extension of Lazarevic's findings that this thesis paper is undertaking. The Part 4 Dataset includes the timeframe from the 1st of September 2017 to the 23rd of September 2020. The Part 5 Dataset includes the timeframe from the 6th of June 2014 to the 23rd of September 2020.

Each of these subsections of the entire dataset have different characteristics that showcase the progressive downward shift of the entire yield curve. As it can be seen in Table 4.2, the progressive shift into negative interest rates reached its peak with the Part 4 Dataset. This is the case because in the Part 4 Dataset the 10y maturity minimum touched a yield of -0.82%, the maximum touched a yield of 0.84% and the mean amounted to a yield of 0.13%. Meaning that if an investor bought a 10y AAA-rated bond within the euro area between 2017 and 2020, she or he would on average receive only 0.13% of yield on her or his investment in 10 years from now. An investment performed between 2014 to 2017 on a 10y AAA-rated bond within the euro area, would on average instead yield 0.52%. Had the same investment been performed between 2004 and 2007, it would have on average yielded 3.81%. The average return of the investment on a 1y maturity AAA-rated bond dropped from an average of 3.04% between 2004 to 2007, to an average of -0.44% between 2014 to 2017, to an average of -0.69% between 2017 and 2020. The current deposit rate of the ECB is -0.50% - this means for example that a bank would actually lose on average 19 basis points (bps) of money if it decided to invest in the 1y bonds markets instead of simply depositing its liquidity at the Central Bank for one year of time between 2019 and 2020[7].

Being that all the min-max ranges across all the maturities are positive in the Part 2 Dataset - this is the dataset considered to be representative of a positive interest rate environment. The min-max ranges across the curve in the Part 3 Dataset have negative values in the min values up to the 10y maturities. The is even more negative in the Part 4 Dataset. In this, the min-max ranges are completely negative from the 1y to the 4y maturities. Additionally, in the Part 4 Dataset, in the min-max range of the very end of the curve, the 30y bond, has a min value touching a negative value of -0.43% - hinting that some european yield curves within the timeframe of 2017 to 2020 went actually entirely negative and entirely below zero[8].

From the Figures 4.2, 4.3 and 4.4 respectively, it is possible to see the path of the entire European term structure from 2004 to 2020. In the first one, it is for example possible to notice the inversion of the curve between 2004 and 2007. The curve from 2007 is relatively flatter than the curve of 2004, which signals higher uncertainty in the short term compared to the long term - this takes place usually in times of

[7]The deposit rate moved to -0.50% in 2019, see Table 4.3.

[8]This occurrence started for the first time in August 2019 with the **German Yield curve**.

Figure 4.2: Yield curves of AAA-rated bonds in the European area from 2004 to 2020

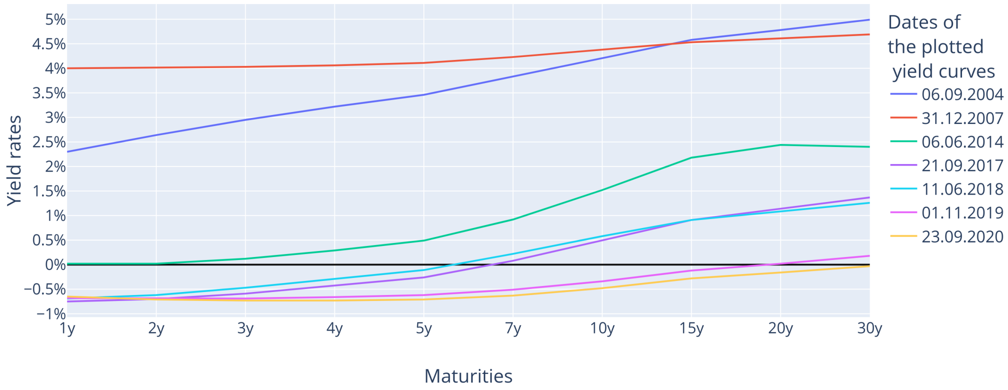


Figure 4.3: Bond Maturities over time, from 2004 to 2020

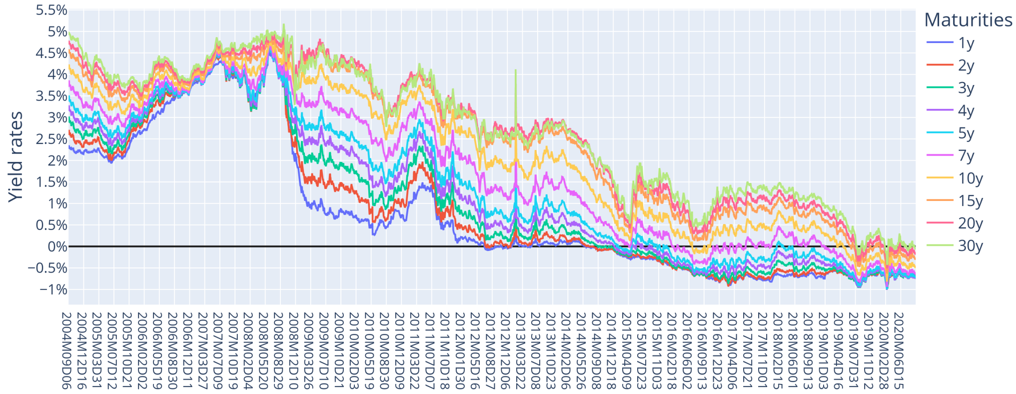
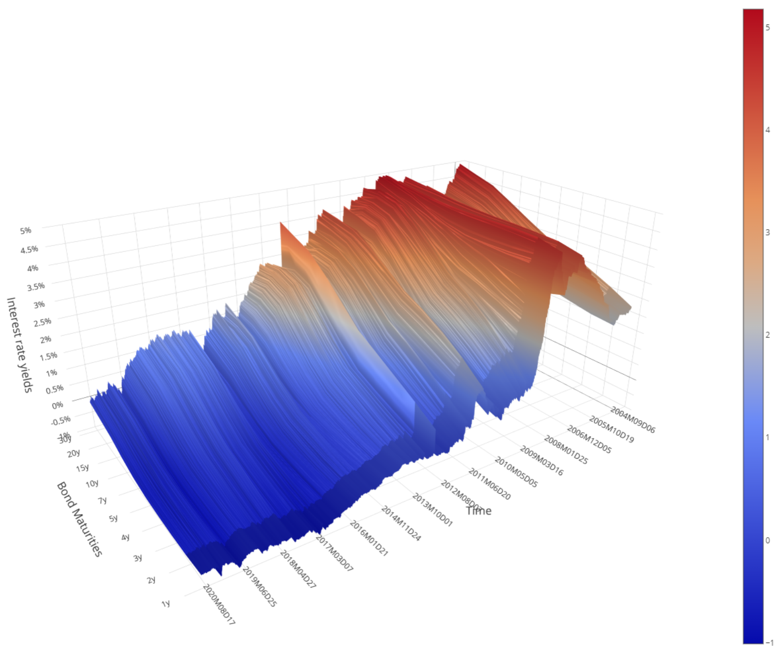


Figure 4.4: Daily downward shift of the yield curve from 2004 to 2020



crisis, as it was at the time. In this same graph there is the depiction of the several downward shifts that the yield curve of the average of the AAA-rated bond had in the last 15 years. In the Figure 4.3 the time series of the individual maturities is illustrated and shown to have moved into the negative space for all maturities. Both of these figures can be summarized with the surface graph in 4.4, where all yield curve shifts are connected in the depicted surface. The more blue the surface, the more near to 0 and to a negative interest rate space.

4.3 ECB's policies: From positive to negative interest rates

As I am writing this paper, since 2014, the euro area finds itself in the presence of negative interest rates. This finds its explanation in the monetary policies that the European Central Bank (ECB) put into place since the global financial crisis of 2007-2008 and the European debt crisis that started at the end of 2009. It was at the height of the European debt crisis and at the height of the spread of several European bonds to the German Bund, that one of the most famous sentences from a central banker marked the history of European and global economics. On the 26th of July 2012 Mario Draghi spoke the famous words "Whatever it takes" in a speech where markets started a path to regaining confidence in the euro area. To restore confidence in the Euro, confidence in European debt had to be restored - this has been possible through a series of monetary policy decisions that tackled the Deposit Facility Rate, access to liquidity, and open market operations that could consequently tackle and lower the yields of European public debt markets.

Figure 4.5: The monthly target of the APP and the cumulative net purchases of the APP by programme [9]

The channels of action that the **ECB** can undertake include setting the key interest rates (such as the deposit interest rate) as well as quantitative easing measures (or 'non-standard monetary policy measure'). In Table 4.3 the rate cuts and hikes that the ECB performed since 2002 up to now are to be found - noting that the deposit facility rate started to be set as negative since 2014. Additionally, in order to ensure liquidity in the markets and to prevent interest rates to spike up again, the ECB started in 2014 to implement a 'non-standard monetary policy measure' of quantitative easing, called the APP, the asset purchase programmes. The

[9]Source: **ECB**.

APP doesn't involve only government bonds (the public sector purchase programme, PSPP), but also corporate bonds (corporate sector purchase programme, CSPP), as well as asset-backed securities (ABSPP) and covered bonds (CBPP3). Figure 4.5 shows the amount of APP in cumulative terms since 2015 for the net purchases - the amount of liquidity that this programme contributed to the markets is of the significant sum of about 3000 billion euros.

As it can be seen from Table 4.3, the ECB decided to cut rates gradually from 2012 to 2019 from 0 to negative fifty basis points - this accommodating policy was chosen in order to provide a great deal of more support for the markets and the economy, with the intention to push the transmission mechanism by helping banks have cheaper access to liquidity and ensuring this access to liquidity was given to consumers, who then would spend, and so on. However, an over-accommodating monetary policy can come with frictions as well. Due to regulatory requirements, European banks need to have a set amount of reserves deposited at the ECB - this compulsory requirement comes at a cost when banks have to pay to deposit these reserve requirements. This has seemed to be a bottleneck for the aforementioned transmission mechanism to function. Given this, in September 2019 a **two-tier system** was introduced, where European banks can have a part of their deposited excess reserves, under some given conditions, at a 0% interest rate instead of the disadvantageous -0.50% rate.

Table 4.3: ECB Deposit Facility Rate from 2004 to 2020

Date		*Deposit Facility*
2019	18 Sep.	-0.50
2016	16 Mar.	-0.40
2015	9 Dec.	-0.30
2014	10 Sep.	-0.20
	11 Jun.	-0.10
2012	11 Jul.	0.00
2011	14 Dec.	0.25
	9 Nov.	0.50
	13 Jul.	0.75
	13 Apr.	0.50
2009	8 Apr.	0.25
	11 Mar.	0.50
	21 Jan.	1.00
2008	10 Dec.	2.00
	12 Nov.	2.75
	9 Oct.	3.25
	8 Oct.	2.75
	9 Jul.	3.25
2007	13 Jun.	3.00
	14 Mar.	2.75
2006	13 Dec.	2.50
	11 Oct.	2.25
	9 Aug.	2.00
	15 Jun.	1.75
	8 Mar.	1.50
2005	6 Dec.	1.25
2003	6 Jun.	1.00
	7 Mar.	1.50
2002	6 Dec.	1.75

Source: **ECB**.

After this, in March 2020 the global pandemic caused by the Covid-19 (known as Corona Virus) hit Europe and the US alike as well as the entire world - this caused a panic in the markets that affected the entire stock markets and bond markets alike. The prompt intervention of the ECB has been the Pandemic Emergency Purchase Programme (**PEPP**) - the entire envelope of this programme is of 1.350 billion euros to date. Not only this, but in light of the pandemic the ECB has also generously set an exclusively accommodating interest rate for the TLTRO III[10] operations between June 2020 and June 2021. In this time-frame, banks getting loans from the **TLTRO III** window will not be borrowing at -0.50% but rather at 50 basis points cheaper, i.e. at -1.0%. Furthermore, there is the expectation that an accommodating monetary policy will continue not only due the pandemic and the demand and supply shock that this caused, but also due to the fact that the ECB has been struggling to fulfill its mandate (an **inflation** close but below 2%).

All the monetary policy measures that have been mentioned here ensure a level of liquidity in the markets that keeps the front end and most of the yield curve anchored negatively. This can be seen also from the downward shift of the entire yield curve that took place between 2014 and 2020 - see aforementioned Figure 4.4 for reference.

[10]The first series of targeted longer-term refinancing operations was announced in June 2014. The second in march 2016, and the third in March 2019. The TLTROs are another instrument, like the APP and PEPP, that the ECB can use in order to preserve favourable borrowing conditions and stimulate bank lending to the real economy.

Chapter 5

Empirical Results

5.1 Shift, Slope and Curvature

In this section, the paper will demonstrate what was presented in the section 2.1. As exposed in Chapter 2.1, previous literature affirms that the first three principal components represent the shift, slope and curvature of the yield curve and that these three components alone can explain the vast majority of the yield curve variability. This is shown empirically to be the case also from the correlation between the PC1 and the shift, PC2 and the slope and PC3 and the curvature. The dataset used in this section is the Part 1 Dataset.

The fact that the first 3 PCs represent the vast majority of the yield curve variability can be shown by the Table 5.1. The first 3 PCs alone can explain together already more than 90% of the total variance. As shown in the table below, the first 3 PCs are representing 92.353% of the total variability. This is a satisfactory percentage given that the PCA calculation led to a total of 10 principal components.

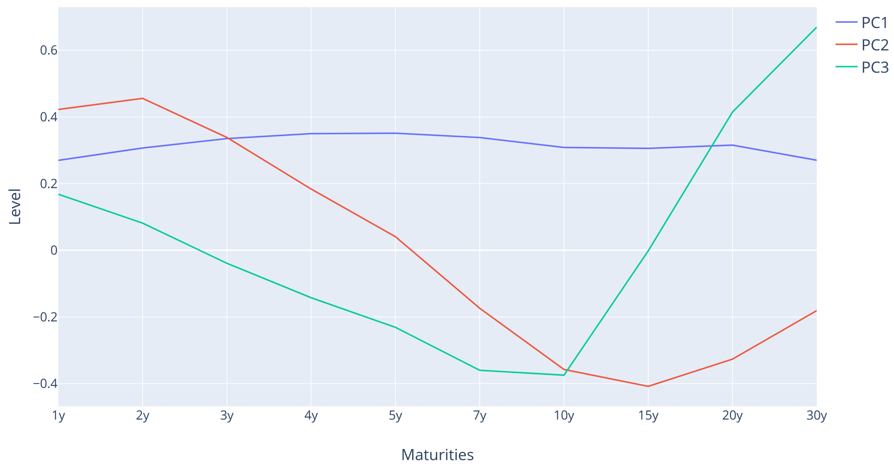
Table 5.1: PCA results from Part 1 Dataset

Variables	$PC1$	$PC2$	$PC3$	$PC4$
$1y$	0.26967	0.422346	0.1681428	0.56076
$2y$	0.30679	0.455693	0.0814210	0.19191
$3y$	0.33502	0.338424	-0.0390552	-0.16647
$4y$	0.34984	0.183643	-0.1423547	-0.32499
$5y$	0.35111	0.040809	-0.2310796	-0.34277
$7y$	0.33817	-0.173813	-0.3599548	-0.16261
$10y$	0.30842	-0.357454	-0.3746351	0.25289
$15y$	0.30561	-0.408218	-0.0018212	0.41612
$20y$	0.31537	-0.326600	0.4145866	0.08241
$30y$	0.27005	-0.180976	0.6694075	-0.35715
Eigenvalues	7.15358	1.29855	0.78319	0.53229
Proportion of Variance	71.54%	12.98%	7.832%	5.323%
Cumulative Variance	71.54%	84.52%	92.353%	97.676%

Next, in order to demonstrate these correlations, we need to identify the concepts of shift, slope and curvature with actual data coming from the yield curve. Shift, slope and curvature can visually be identified with the 10y maturity, the 2y10y spread and the 7y15y30y butterfly spread.

The concept of shift can be visualised with the 10y maturity. This is because with the term shift it is meant the shift of the entire yield curve. If the front end (the front end part of the curve means the short term maturities) and the long end

Figure 5.1: Plotting of the first 3 PCs from the Part 1 Dataset - Results from Table 5.1



(the longer term maturities) shift lower and lower, the 10y maturity and the rest of the curve will shift accordingly. Being the 10 year maturity here in the "middle" of the curve and one of the most liquid points out of the yield curve, this will be the maturity chosen to represent the shift of the yield curve.

The concept of slope can be visualised with the 2y10y spread. A spread is the difference between two financial assets, here bonds and their yields. The spread considered for the slope is the common 2y10y spread. This is calculated by the 10 years yield minus the 2 years yield. The term slope has been referred to in previous literature also as steepness. With this kind of spread, "steepener" or "flattener" trades are possible. A "steepener" trade would be a trade where the front end is bought and the long end is sold, because the bet is that the curve would steepen. A "flattener" trade is a trade where the front end is sold and the long end is bought, because the bet is that the curve would flatten or invert.

The concept of curvature can be visualised with the 7y15y30y "butterfly" spread. A butterfly spread in financial terms is the spread within a butterfly trade. A butterfly trade consists of picking two points on the curve and selling (buying) them while simultaneously picking a point between these two and buying (selling) it. The former trade is performed if one is betting that the curve in the chosen part of the curve's curvature is going to look more convex. The latter is performed if one is betting that the chosen part of the curve's curvature is going to look more concave. The butterfly trade can be expressed in any part of the curve – but the butterfly spread chosen here is to express the curvature of the entire curve, hence the picked butterfly spread here is the 7y15y30y spread. This is calculated by multiplying the 15 years yield by two and subtracting the 7 years yield and the 30 years yield.

Table 5.2: Correlations in absolute value of PC1, PC2, PC3 with SSC

	PC1 and 10y	*PC2 and 2y10y spread*	*PC3 and 7y15y30y Butterfly spread*
Corr	0.81090	0.82205	0.65252

As it can be seen from Table 5.2, the correlations in absolute value between these pairs is higher than 0.5 - which illustrates how the PC1 represents the shift, the PC2 represents the slope and the PC3 represents the curvature of the yield curve.

Figure 5.2: First three PCs with 10y bond, 2y10y spread and 7y15y30y spread[1]

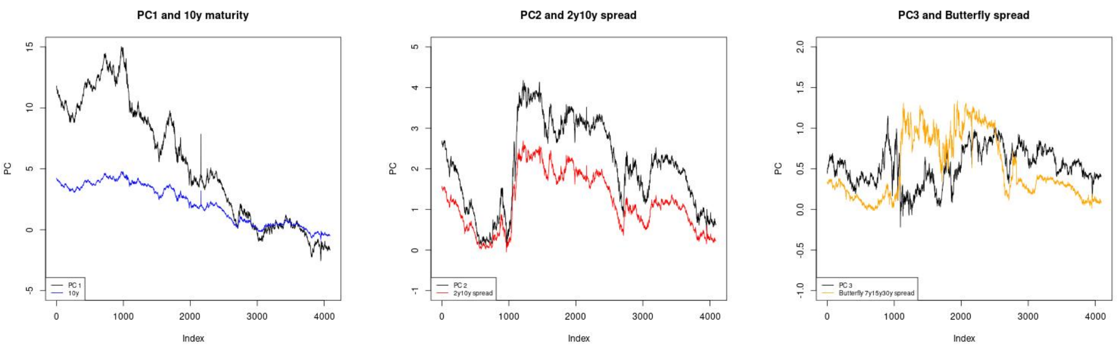
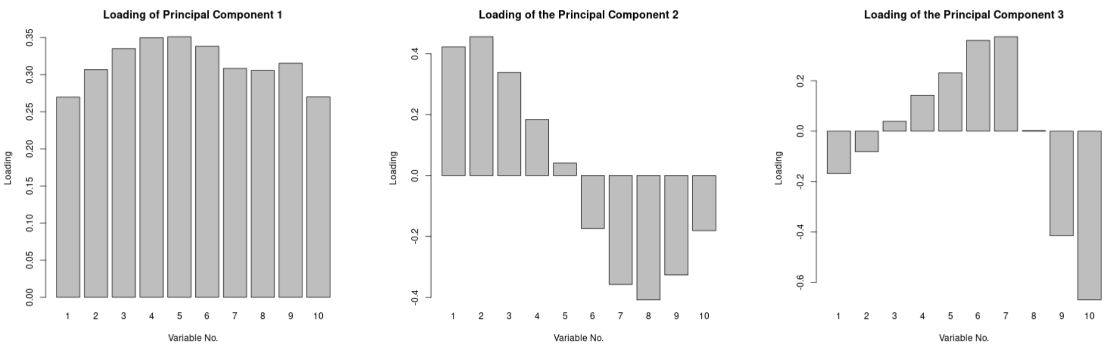


Figure 5.3: First three PCs loadings[2]



Additionally, this can also be noted visually from Figure 5.2. Lastly, also here it's implied that the sign changes can be an additional characteristic towards the SSC pattern. As it can be seen from Figure 5.3, the loading of the PC1 shows no sign changes while the loading of PC2 leads to 1 sign change and the loading of PC3 to 2 sign changes. This also is a hint towards the shift, slope and curvature associations.

[1] These graphs come from the principal component analysis applied on the entire dataset without taking the first difference. Although stationarity is not ensured in this case, these graphs are more communicative for illustration purposes to depict the correlation between the aforementioned pairs. In the Appendix B.2 there are the graphs of the analysis from the PCA applied to the actual Part 1 Dataset - which has the first differences taken, hence with ensured stationarity.

[2] These graphs come from the principal component analysis applied on the entire dataset with the first difference taken, hence with ensured stationarity.

5.2 Positive Interest Rates Environment vs Negative Interest Rates Environment

This section will first tackle the replication of Lazarevic's work from his 2019 paper "Principal Component Analysis in Negative Interest Rate Environment". This first replication aims to show what was discussed in Chapter 2.2. The replication of Lazarevic's work is performed with Part 2 and 3 Datasets. After this, the application of PCA to the average of AAA-rated bonds of the european area will be extended to the time frame of 2017-2020 with the Part 4 Dataset and to the timeframe of 2014-2020 with the Part 5 Dataset.

Replication of Lazarevic (2019) results

As it can be seen from Table 5.3 and Table 5.4 - the results are successfully replicated and meet the expectations set by Lazarevic (2019). Within the positive interest rate environment between 2004 and 2007, the cumulative explained variance from the first 3 PCs is higher than 95%, amounting to 97.265%. As expected and as shown previously by Lazarevic, the Part 2 Dataset does not require the consideration of an additional PC to reach a cumulative explained variability of 95%. The 'oscillatority' term, the fourth PC, is hereby not necessary. Lazarevic refers to the 4th PC as the 'oscillatority' given its 3 sign changes.

Table 5.3: PCA results from Part 2 Dataset - Positive Interest Rate Environment

Variables	$PC1$	$PC2$	$PC3$	$PC4$
$1y$	0.27491	0.3563383	0.460318	-0.582079
$2y$	0.31107	0.3295822	0.355936	0.037902
$3y$	0.33501	0.3031091	0.086143	0.339697
$4y$	0.34095	0.2543300	-0.180909	0.335170
$5y$	0.33715	0.1773714	-0.370480	0.190018
$7y$	0.33061	-0.0026005	-0.491468	-0.220174
$10y$	0.33225	-0.2255638	-0.298368	-0.456511
$15y$	0.31298	-0.3978525	0.084421	-0.150122
$20y$	0.29481	-0.4277243	0.242614	0.126800
$30y$	0.28451	-0.4292532	0.295738	0.316884
Eigenvalues	7.18754	1.72461	0.81438	0.16566
Proportion of Variance	71.88%	17.25%	8.144%	1.657%
Cumulative Variance	71.88%	89.12%	**97.265%**	98.922%

Table 5.4: PCA results from Part 3 Dataset

Variables	$PC1$	$PC2$	$PC3$	$PC4$
$1y$	0.20255	-0.5042000	-0.532377	-0.458625
$2y$	0.27687	-0.5110871	-0.028416	-0.016598
$3y$	0.32635	-0.3293045	0.208593	0.246288
$4y$	0.35167	-0.1557367	0.226657	0.296576
$5y$	0.35955	-0.0073399	0.191564	0.243467
$7y$	0.358222	0.1464340	0.148080	0.017921
$10y$	0.34398	0.2443458	0.166538	-0.319965
$15y$	0.33001	0.2900994	0.087689	-0.439257
$20y$	0.32831	0.3286593	-0.216936	-0.155194
$30y$	0.24400	0.2801799	-0.692657	0.511630
Eigenvalues	7.08194	1.48939	0.73575	0.40496
Proportion of Variance	70.82%	14.89%	7.357%	4.05%
Cumulative Variance	70.82%	85.71%	**93.071%**	**97.12%**

As expected and shown previously by Lazarevic (2019), the Part 3 Dataset does instead require an additional PC to reach an as satisfactory amount of cumulative explained variance. The first three PCs explain only the 93.071% of total variability. In order to achieve at least 95% the 'oscillatority' term, the fourth PC, should in this case be taken into consideration.

Extension of Lazarevic (2019) results

The same calculations performed in R that led to the results from Tables 5.3 and 5.4 have been applied to the Part 4 and 5 Datasets to produce the results displayed in the Tables 5.5 and 5.6.

As shown in Chapter 4, in the Part 4 Dataset (timeframe of 2017 to 2020) there lies an even deeper negative interest rates environment compared to the Part 3 Dataset (timeframe of 2014 to 2017). This can be numerically seen by comparing the lower mean and lower min values that the Part 4 Dataset entails compared to the Part 3 Datasets in the Table 4.2. Furthermore, this can also be visually seen in Figure 4.4: the green yield curve from 2014 and the violet yield curve from 2017 lie above the fuchsia and yellow yield curves from 2019 and 2020 respectively.

As a consequence, given Lazarevic (2019) argumentation, there is the expectation that the results from Part 4 and 5 Datasets also lead to a cumulative explained variance of the first three PCs to be lower than 95%. This has not been found to be the case.

Table 5.5: PCA results from Part 4 Dataset

Variables	$PC1$	$PC2$	$PC3$	$PC4$
$1y$	0.24607	0.42116	-0.67312	0.39692
$2y$	0.30348	0.40775	-0.20302	-0.26325
$3y$	0.32474	0.31478	0.13716	-0.36450
$4y$	0.33335	0.22051	0.32235	-0.18603
$5y$	0.33946	0.12558	0.37802	0.07144
$7y$	0.34212	-0.04953	0.31829	0.39065
$10y$	0.33624	-0.22651	0.08085	0.47839
$15y$	0.32003	-0.35240	-0.15144	0.08210
$20y$	0.31022	-0.38824	-0.21516	-0.18874
$30y$	0.29461	-0.40526	-0.25097	-0.420347
Eigenvalues	7.708571	1.41973	0.541584	0.190582
Proportion of Variance	77.09%	14.20%	5.416%	1.906%
Cumulative Variance	77.09%	91.28%	**96.699%**	98.605%

Table 5.6: PCA results from Part 5 Dataset

Variables	$PC1$	$PC2$	$PC3$	$PC4$
$1y$	0.22483	-0.47526	-0.57485	0.38614
$2y$	0.29018	-0.47185	-0.10984	0.01369
$3y$	0.32697	-0.32616	0.17689	-0.20654
$4y$	0.34565	-0.17936	0.27108	-0.26404
$5y$	0.35331	-0.04515	0.27529	-0.21551
$7y$	0.35258	0.12336	0.22546	0.00124
$10y$	0.33948	0.24611	0.15667	0.34876
$15y$	0.32447	0.31354	0.01507	0.44015
$20y$	0.32082	0.35559	-0.22783	0.11709
$30y$	0.25742	0.33221	-0.59110	-0.60234
Eigenvalues	7.27230	1.47897	0.63724	0.35794
Proportion of Variance	72.72%	14.79%	6.372%	3.579%
Cumulative Variance	72.72%	87.51%	**93.885%**	**97.464%**

The cumulative explained variance of the first three PCs from the Part 4 Dataset amounts to almost 97%. This leads to the needlessness of taking also the oscillatority factor into consideration. This finding does not meet the expectation - but further analysis has been pursued to consider the entirety of the negative interest rate environment dataset since its start (2014 to 2020). Hence, the same calculations have been applied to the Part 5 Dataset and the findings are interesting as they contradict the findings resulting from the 4 Dataset.

In Table 5.6, it can be seen that the cumulative explained variance of the first three PCs does not meet 95% but 93.885% instead. In this case, the fourth PC does have to be taken into consideration to ensure a higher proportion of explained variance, to meet the same efficacy of PCA applied to the positive interest rates environment.

Table 5.7: Explained Variances in Comparison

	$PC1$	$PC2$	$PC3$	$PC4$
Results from Part 2 Dataset - the positive interest rate environment				
Eigenvalues	7.18754	1.72461	0.81438	0.16566
Proportion of Variance	71.88%	17.25%	8.144%	1.657%
Cumulative Variance	71.88%	89.12%	**97.265%**	98.922%
Results from Part 3 Dataset - the negative interest rate environment				
Eigenvalues	7.08194	1.48939	0.73575	0.40496
Proportion of Variance	70.82%	14.89%	7.357%	4.05%
Cumulative Variance	70.82%	85.71%	**93.071%**	**97.12%**
Results from the Extension Part 4 Dataset - negative interest rate environment				
Eigenvalues	7.708571	1.41973	0.541584	0.190582
Proportion of Variance	77.09%	14.20%	5.416%	1.906%
Cumulative Variance	77.09%	91.28%	**96.699%**	98.605%
Results from the Extension Part 5 Dataset - negative interest rate environment				
Eigenvalues	7.27230	1.47897	0.63724	0.35794
Proportion of Variance	72.72%	14.79%	6.372%	3.579%
Cumulative Variance	72.72%	87.51%	**93.885%**	**97.464%**

As shown, the results from the Part 4 Dataset do not meet the expectations but the results from the 5 Dataset do. This is rather contradictory given the fact that on average the Part 5 Dataset is less negative than the Part 4 Dataset - this can be seen numerically by comparing the means in Table 4.2. As a consequence, this leads to the question whether it is the negativity of interest rates itself that determines the lower explained variability seen with the Part 3 Dataset. Had the negativity of interest rates been the cause for this, a lower amount of variability would have applied also to the results from the Part 4 Dataset. As it did not apply, this leads to the suggestion that there are some idiosyncratic characteristics in the Part 3 Dataset, other than the negativity of interest rates, that influence a lower explained variability of the first three PCs.

Chapter 6

Conclusion

PCA is a dimensionality reduction technique that can be applied to the term structure to describe an entire interest rates market with only a few factors, while still explaining the majority of the variability of the original dataset. From previous literature these hae been considered to be the shift for the first PC, the slope for the second PC and the curvature for the third PC (SSC).

This research is able to empirically show that the SSC holds and that it be can graphically modelled as a long end rate (here the 10y maturity for the first PC), the 2y10y spread (for the slope, i.e. the second PC) and the 7y15y30y butterfly spread (for the curvature, i.e. the third PC).

Previous literature (Lazarevic, 2019) researched about whether the SSC pattern can still be found in a negative interest rate environment. Though this has been found to be the case, the amount of variability explained by the first three PCs seems to have been unfavourably impacted by the presence of negative interest rates. With regards to the efficacy of PCA applied to the term structure, given the previous literature (Lazarevic, 2019), the expectation was that the Part 4 and 5 Datasets would result in a similar outcome as the previous literature showed. Meaning that the amount of variability explained by the first three PCs in both of these Datasets would be lower than the amount of explained variability of the first three PCs from a positive interest rates sample (here, Part 2 Dataset).

However, this expectation has not been met from the analysis resulting from the Part 4 Dataset. Contrarily, this has been found to be the case for the Part 5 Dataset. The fact that a lower (in respect to the positive interest rate environment in the Part 2 Dataset) explained variability of the first 3 PCs results from the Part 3 Dataset and not the Part 4, makes this a contradictory outcome. This contradictory outcome leads and results in questioning whether the negativity of interest rates is the actual cause for a lower explained variability of the first 3 PCs resulting from the Part 3 Dataset, which instead suggests an idiosyncrasy within the Part 3 Dataset.

The conclusion of this research regarding the efficacy of PCA for the term structure in a negative interest rate environment is that the negativity of interest rates might actually not affect the amount of variability explained by the first three PCs. Although the research's results from the previous literature indicates this to be the case, here this is not. The efficacy in a very deeply negative interest rate environment (in Part 4 Dataset the cumulative explained variance from the first three PCs is 96.69%) is found to be as high as the efficacy of PCA in a positive interest rate environment (in Part 2 Dataset the cumulative explained variance from the first three PCs is 97.27%). This suggests that the presence of other idiosyncratic characteristics of a specific time series might affect the efficacy of PCA as an instrument applied to the term structure.

Bibliography

Fabozzi, F. J. (1983). *The handbook of fixed income securities.* McGraw-Hill, 7th Edition.

Huggins, D., & Schaller, C. (2013). *Fixed income relative value analysis.* Wiley, Bloomberg Press.

Hull, J. C. (2015). *Options, futures, and other derivatives.* Pearson, 9th Edition.

Joliffe, I. T. (2002). *Principal component analysis.* Springer, 2nd Edition.

Lazarevic, M. (2019). Principal component analysis in negative interest rate environment. *Acta Oeconomica*, *69*(1), 101–125.

Lekkos, I. (2000). A critique of factor analysis of interest rates. *The journal of derivatives*, *8*(1), 72-83.

Litterman, R., & Scheinkman, J. (1991). Common factors affecting bond returns. *The journal of fixed income*.

Lord, R., & Pelsser, A. (2006). Level-slope-curvature - fact or artefact? *Applied Mathematical Finance*, *14*(2), 105–130.

Ponomareva, N., Sheen, J., & Wang, B. (2019). The common factor of bilateral u.s. exchange rates: What is it related to? *Empirical Economics*, *56*, 1251–1268.

Rencher, A. C. (2012). *Methods of multivariate analysis.* Wiley - Interscience, 3rd Edition.

Ruppert, D., & Matteson, D. S. (2015). *Statistics and data analysis for financial engineering.* Springer.

Salinelli, E., & Sgarra, C. (2006). Correlation matrices of yields and total positivity. *ScienceDirect*, *418*, 682–692.

Salinelli, E., & Sgarra, C. (2007). Shift, slope and curvature for a class of yields correlation matrices. *ScienceDirect*, *426*, 650–666.

Steeley, J. (1990). Modelling the dynamics of the term structure of interest rates. *The Econoomic and Social Review*, *21*(4), 337-361.

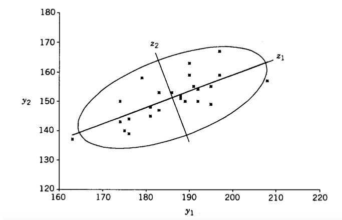
Appendix A

PCA

A.1 Principal Components Derivation

The main idea of principal component analysis is to reduce the dimensionality of the dataset forming new combinations of the existing variables. The first PC is the linear combination with maximal variance. The second PC is the linear combination with maximal variance in a direction orthogonal (90 degrees) to the first principal component, and so on (Rencher, 2012). The orthogonality of the eigenvectors is an important assumption, because it makes it possible to decompose the covariance matrix into uncorrelated relationships.

Figure A.1: Orthogonal vectors[1]



The following proofs have been selected from Joliffe (2002). A $n \times k$ Matrix X and a $k \times 1$ vector α_i are given. A Principal component can come into place as a linear combination of all the features where the weights are given by the elements of α_i. The k_{th} principal component (PC) is given by

$$z_k = \alpha'_k \text{ x} = \alpha'_{k1}\ x_2 + \alpha'_{k3}\ x_3 + \ldots + \alpha'_{kp}\ \text{x}_p = \textstyle\sum_{j=1}^{p} \alpha'_{kj}\ \text{x}_j$$

Given that the first principal component has to be the one with the highest variance, the first PC comes into place from an optimization problem:

$$\max: \quad var(z_k) = var(\alpha'_k x) \qquad s.t. \quad \alpha'_1\alpha_1 = 1$$

The maximization of the first PC presents itself as:

$$\max: \quad var(z_1) = var(\alpha'_1 x) \qquad s.t. \quad \alpha'_1\alpha_1 = 1$$

[1]Source of this picture: Rencher (2012, page 384).

Which, given that Σ is the covariance matrix of X, is equivalent to:

$$\max : \alpha_1' \Sigma \alpha_1 \qquad s.t. \quad \alpha_1' \alpha_1 = 1$$

From this, it follows a Lagrange:

$$\begin{gathered} L = \alpha_1' \sum \alpha_1 - \lambda \left(\alpha_1' \alpha_1 - 1 \right) \\ \frac{\partial L}{\partial \alpha_1} = \sum \alpha_1 - \lambda \alpha_1 = 0 \\ \left(\sum - \lambda I \right) \alpha_1 = 0 \\ \sum \alpha_1 = \lambda \alpha_1 \mid \cdot \alpha_1' \\ \alpha_1' \sum \alpha_1 = \alpha_1' \lambda \alpha_1 = \lambda \alpha_1' \alpha_1 = \lambda_1 \end{gathered}$$

To decide which of the eigenvectors p gives the max variance of $z_k = \alpha_k' x$, the quantity to be maximised is $\sum \alpha_1$, so λ_1 must be as large as possible.

Also the problem of the second PC is a maximisation problem, just with an additional optimization constraint: that the the second PC is uncorrelated with the first PC, otherwise they would entail parts of the same information baggage, which is not what we aim for in PCA. Orthogonality $\alpha_1' \alpha_2 = \alpha_2' \alpha_1 = 0$ here is actually mathematically also necessary to ensure that the covariance between the first and the second eigenvectors is 0 - hence to ensure that these are uncorrelated also means to ensure orthogonality. The maximization of the second PC presents itself as:

$$\max : \alpha_2' \Sigma \alpha_2 \qquad s.t. \quad cov\left[\alpha_1' x, \alpha_2' x\right] = 0 \quad \& \quad s.t. \quad \alpha_2' \alpha_2 = 1$$

However, we know that in order for the following to hold:

$$cov\left[\alpha_1' x, \alpha_2' x\right] = \alpha_1' \Sigma \alpha_2 = \alpha_2' \Sigma \alpha_1 = \alpha_2' \lambda \alpha_1' = \lambda_1 \alpha_2' \alpha_1 = \lambda_1 \alpha_1' \alpha_2 = 0$$

...it has to hold also orthogonality so that: $\alpha_1' \alpha_2 = \alpha_2' \alpha_1 = 0$. Which turns our optimization problem for the second PC into:

$$\begin{gathered} \max : \alpha_2' \sum \alpha_2 \qquad s.t. \quad \alpha_2' \alpha_2 = 1, \alpha_2' \alpha_1 = 0 \\ L = \alpha_2' \sum \alpha_2 - \lambda \left(\alpha_2' \alpha_2 - 1 \right) - \phi \left(\alpha_2' \alpha_1 - 0 \right) \\ \frac{\partial L}{\partial \alpha_2} = \sum \alpha_2 - \lambda \alpha_2 - \phi \alpha_1 = 0 \end{gathered}$$

From this, to understand what value ϕ has, we multiply on the left for α_1':

$$\begin{gathered} \frac{\partial L}{\partial \alpha_2} = \sum \alpha_2 - \lambda \alpha_2 - \phi \alpha_1 = 0 \mid \cdot \alpha_1' \\ \alpha_1' \sum \alpha_2 - \lambda \alpha_1' \alpha_2 - \alpha_1' \phi \alpha_1 = 0 \end{gathered}$$

Here we know from above the first two terms to be zero - hence the third term must be zero. For the third term to be zero, ϕ has to be zero. Which leads to:

$$\begin{gathered} \sum \alpha_2 - \lambda \alpha_2 = 0 \\ \sum \alpha_2 = \lambda \alpha_2 \\ (\Sigma - \lambda I) \alpha_2 = 0 \\ \lambda_2 = \alpha_2' \sum \alpha_2 \end{gathered}$$

This means that λ_2 has to be as large as possible. Assuming that Σ does not have repeated eigenvalues, then λ_2 cannot equal λ_1. If it did, then also $\alpha_2 = \alpha_1$, which would violate $\alpha_1' \alpha_2 = 0$. This means that the second PC has the second largest variance.

As a consequence, α_k is an eigenvector of the covariance matrix Σ corresponding to its k_{th} largest eigenvectors. Also, α_k is chosen to have unit length of $\alpha_k' \ \alpha_k = 1$ so that var(z_k) $= \lambda_k$.

This is the case because the variances of the PCs are related to the eigenvalues. This can be seen from the following.

The variance-covariance Matrix Σ is always symmetric and always semi-positive definit. The properties that unfold from these charachteristics are that: all eigenvalues are positive, the eigenvectors are orthogonal (uncorrelated), there is a full set of eigenvalues with corresponding eigenvectors (meaning that there are as many PCs as variables in the dataset X).

Given these, the covariance matrix can be 'broken' into three other parts. This is called spectral decomposition. The first part, V, is a $k \times k$ matrix with all the eigenvectors as columns, the second part Λ is a diagonal $k \times k$ matrix with all eigenvalues as diagonal elements and the third part is the matrix V transposed:

$$\Sigma = cov(X) = V\Lambda V'$$

Furthermore, $\Sigma = V\Lambda V'$ can be rewritten as $V'\Sigma V = (V'V)\Lambda(V'V) = I \wedge I = \Lambda$ and the PC, Γ, is given by $\Gamma = XV$. The PCs that result from this are in the $n \times k$ matrix Γ, as X is a $n \times k$ and V is a $k \times k$ matrix. As stated above, there are always as many PCs as variables in the dataset X.

It follows that the variance of the PCs are related to the eigevalues:

$$var(\Gamma) = var(XV) = V'\Sigma V = \Lambda$$

Appendix B

Graphs

B.1 ACF Graph

These are the Autocorrelation Graphs referred to in Chapter 3.1

Figure B.1: Autocorrelation functions within the Part 1 Dataset bond time series - without the first differences taken

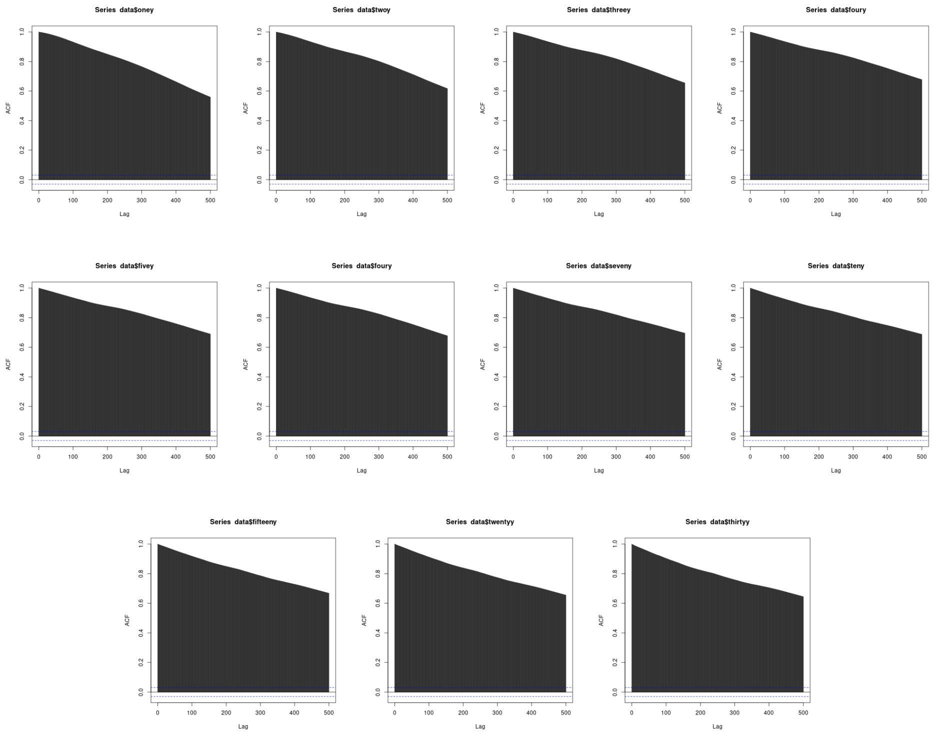
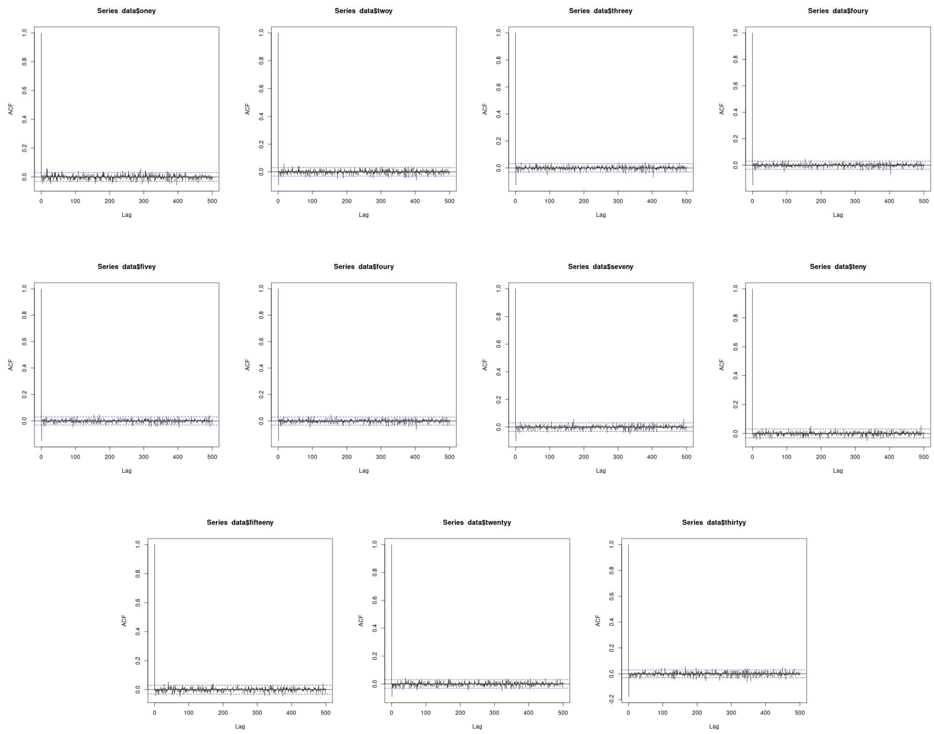


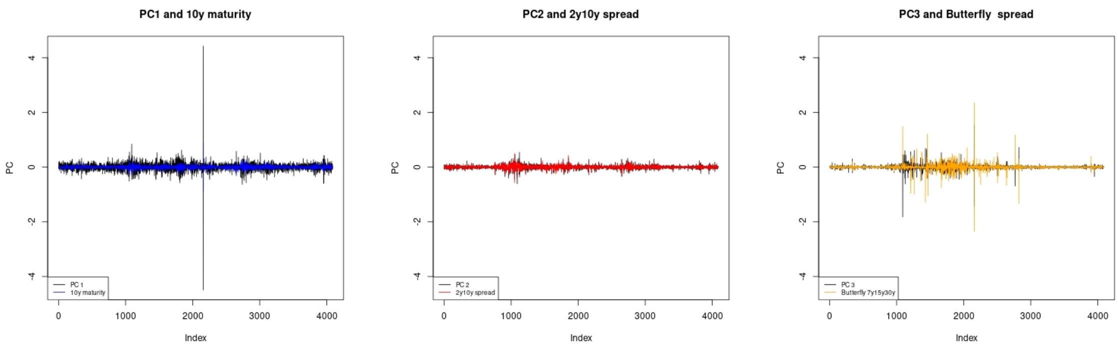
Figure B.2: Autocorrelation functions within the Part 1 Dataset bond time series - with the first differences taken



B.2 SSC graph with stationary dataset

This is the graph referred to in Figure 5.2 These are the graphs that come from calculations from the Part 1 Dataset - which has the first differences taken, hence with ensured stationarity.

Figure B.3: First three PCs with 10y bond, 2y10y spread and 7y15y30y spread



I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Berlin, November 10, 2020

. . . .

(Signature of the author)